ЛЕКЦІЯ 1. ВИЗНАЧЕННЯ ОДНОРІДНОСТІ ПІДМНОЖИН ОБ'ЄКТІВ

НАВЧАЛЬНІ ПИТАННЯ

Види однорідних підмножин

Метод куль

1. Види однорідних підмножин

Реалізації багатомірних величин є об'єктами таксономічних досліджень, тому їх розглядають, як вектори , або точки, які розташовані в багатомірному просторі. Ці об'єкті розрізняються як по рівню, так і по структурі даних.

Рівень значень визначає довжину вектора , а структура – кут між векторами. Чим більші значення ознак, тим далі від початку координат розташована точка-об'єкт, а отже і тим більша довжина вектору. З іншого боку, структура показує для кожного об'єкту частку значень кожної ознаки в їх загальній сумі. Отже, однакові структури будуть спостерігатись для колінеарних векторів, в яких пропорції значень однакові.

Вибір однорідності за рівнем, структурою чи обома цими характеристиками, визначається дослідником, виходячи з сутності завдання, що розв'язується і цілей дослідження. Відповідно, розрізняють такі види однорідності:

- ізотропність – за рівнем та структурою одночасно;
- ізотонічність – за рівнем значень;
- ізоморфічність – за структурою.

Інформація про розташування точок в багатомірному просторі дозволяє прийняти гіпотезу про форму моделі. Якщо точки розташовані у формі еліпсоїда, а розподіл багатомірної випадкової величини відповідає нормальному, то модель має лінійний вигляд.

Однорідність даних потрібна для того, щоб правильно відібрати об'єкти для побудови моделі. Однак різні ознаки мають різні одиниці виміру, тому в основі всіх методів лежить узгодження даних. Наприклад, вартість основних фондів підприємства вимірюється у тис. грн., середньомісячна заробітна плата на одного працюючого – у сотнях грн.., а собівартість одиниці продукції – у грн., або й у копійках. Очевидно, що дисперсії таких ознак, близькі за значеннями, мають різний смисл. Конкретні способи узгодження даних будуть розглянуті пізніше.

Ізотропні підмножини

Найчастіше використовуються такі таксономічні методи, за допомогою яких виділяються підмножини, однорідні в змісті ізотропності (об'єкти, що належать тому самому підмножині, мало відрізняються друг від друга за рівнем і структурою значень ознак). Справа в тім, що, як правило, позбутися від одиниць виміру ознак можна шляхом так званої стандартизації:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \qquad (1.1)$$

де x_{ij} – i-та реалізація j-тої ознаки;

$\bar{x}_j$ – середнє значення j-тої ознаки;

s_j – середньоквадратичне відхилення j-тої ознаки.

У результаті обидві компоненти реалізації ознак (структура і рівень) вирівняні.

Відстань між об'єктами обчислюється за формулою звичайної евклідової метрики

$$d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{kj})^2}. \qquad (1.2)$$

В економетричних дослідженнях подібна однорідність не завжди бажана, оскільки у виділених підмножинах елементи розташовуються в деякій області багатомірного простору властивостей, що має форму кулі. У той же час не можна одержати підмножини, однорідні або за структурою, або за рівнем значень ознак, тобто підмножини, елементи яких розміщаються в просторі властивостей в області, що має форму еліпсоїда. Застосування методів ізотропної таксономії приводить, таким чином, до штучного (якщо говорити про вимоги економетричних методів) розбиття сукупності об'єктів на однорідні підмножини. Такий досить специфічний розподіл сукупності об'єктів на ізотропні підмножини виходить незалежно від того, є чи ні в досліджуваній сукупності об'єктів підмножини, однорідні з погляду структури або рівня значень ознак.

Ізотонічні підмножини

Сукупність операцій, що приводять до одержання підмножин даних, що складаються з елементів, однорідних за рівнем значень ознак, будемо називати ізотонічним аналізом, а отримані підмножини даних - ізотонічними підмножинами.

Перш ніж почати виділення ізотонічних підмножин, потрібно позбутися від одиниць виміру ознак. З цією метою використовуємо перетворення, що виключить зі значень кожної ознаки компонентові структури, але збереже компонентові рівня. Для цього перетворимо ознаки

$X_i = [x_{1i}, x_{2i}, \ldots, x_{ni}]^T$

до виду

$V_i = [v_{1i}, v_{2i}, \ldots, v_{ni}]^T$, (1.3)

за допомогою ізотонічного перетворення

$$v_{ji} = \frac{x_{ji}}{\sum_{j=1}^{m} x_{ji}}. \quad (1.4)$$

Отримані в такий спосіб ознаки характеризуються тим, що їхні довжини (що розуміються як довжини векторів) дорівнюють одиниці, тобто для ознак V_j виконується умова:

$$\|V_i\| = \sum_{j=1}^{m} v_{ji} = 1. \quad (1.5)$$

Вихідні ж ознаки звичайно мають різну довжину.

При одержанні ознак V_j виключаються одиниці виміру вихідних ознак і встановлюються однакові довжини ознак, рівні 1.Таким чином, значення v_{ij} характеризують "потенціал", "масштаб", "позицію" або "ранг" даного об'єкта в розглянутій сукупності об'єктів. У свою чергу сума значень

$$w_j = \sum_{i=1}^{n} v_{ji} \quad (1.6)$$

характеризує внесок значень усіх властивостей даного об'єкта в значення всіх властивостей сукупності, що включає всі об'єктів. Ця величина синтетичним образом представляє "потенціал" або "позицію" *j*-го об'єкта в сукупності. Вектор $W = [w_1, w_2, \ldots, w_m]^T$ називається ізотонічним показником сукупності даних.

Для виділення підмножин, однорідних в ізотонічному відношенні, відстань між точками об’єктами визначається за формулою “міської” метрики:

$$d_{ij} = |w_i - w_j|. \quad (1.7)$$

Елементи d_{ij} приймають значення, рівні нулеві, якщо значення показників w_i і w_j ідентичні.

Узагалі можна затверджувати, що чим вище значення d_{ij}, тим більше відрізняються між собою розглянуті об'єкти за рівнем значень властивостей. Низькі значення d_{ij} свідчать про малі розходження об'єктів за рівнем значень властивостей.

Ізоморфічні підмножини.

Такі підмножини характеризуються однорідністю з точки зору структури значень. Тобто, пропорції значень відповідних ознак будуть мало відрізнятись одна від одної. Для їх одержання потрібно провести наступні перетворення

$$z_{ij} = \frac{\frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}}}{\sum_{j=1}^{n} \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}}} = \frac{v_{ij}}{\sum_{j=1}^{n} v_{ij}}. \quad (1.8)$$

В результаті одержимо нові точки $Ri = (z_{i1}, z_{i2}, \ldots, z_{in})$.Вони характеризуються тим, що $\|R_i\| = \sum_{j=1}^{n} z_{ij} = 1$. Відстань між такими точками визначається також за допомогою звичайної евклідової метрики:

$$d_{ij} = \sqrt{\sum_{k=1}^{m} (z_{ki} - z_{kj})^2}. \quad (1.9)$$

Мінімальна відстань між об’єктами R_i та R_j спостерігається тоді, коли вихідні вектори P_i та P_j колінеарні. Максимум відстані досягається, якщо вихідні вектори ортогональні. Помітимо, що значення перетворених векторів лежать в межах від 0 до 1. Отже в такому випадку максимальна відстань буде становити $\sqrt{2}$.

2. Метод куль

Одним зі способів, який дозволяє здійснити поділ множини об’єктів на однорідні підмножини, є метод куль. Сутність його полягає в тому, що для кожної точки будується куля певного радіуса. Однорідними з даним об’єктом вважаються ті, які потрапляють до даної кулі. При цьому спочатку обирається така куля, яка буде містити найбільшу кількість об’єктів. Далі відібрані об’єкти вилучаються з розгляду, і процедура побудови куль продовжується.

Алгоритм методу має такий вигляд:

1 *крок*. Здійснюється перетворення вихідних даних з метою їх узгодження.

2 *крок*. Для перетворених даних розраховується матриця відстаней *D*.

3 *крок*. Обирається радіус кулі. Це можна зробити за таким правилом:

$$\rho = \max_i \min_{j \neq i} d_{ij}. \quad (1.10)$$

Тобто, для кожного стовпчика матриці відстаней обирається найменший елемент, і в ролі радіуса кулі береться найбільше з цих значень. Зрозуміло, що діагональні елементи матриці відстаней рівні 0, тому виключаються з аналізу даних для пошуку радіуса.

Іншим способом обчислення радіуса кулі є використання формули

$$\rho = \bar{d} + k\sigma_d, \quad (1.11)$$

де

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i, \quad (1.12)$$

$$\sigma_d = \left[\frac{1}{m}\sum_{i=1}^{m}(d_i - \bar{d})^2\right]^{1/2}, \quad (1.13)$$

$$d_j = \min_i d_{ij}, \quad (1.14)$$

k – деяка константа, обирається рівною 2 або 3 (правило 2σ!) .

4 *крок*. Для кожної точки-об’єкта будується куля заданого радіуса і визначається кількість точок, що попали всередину цієї кулі. Технічно це означає підрахунок по матриці відстаней кількості значень, що менші за радіус кулі. Першу однорідну підмножину складуть ті об’єкти, які увійдуть до кулі з найбільшою кількістю елементів.

5 *крок*. Відібрані об’єкти вилучаються з розгляду. Це означає, що з матриці відстаней вилучаються рядки та стовпчики, які відповідають відібраним об’єктам.

6 *крок*. Знову будуються кулі для тих об’єктів, що залишились. Радіус кулі залишається незмінним. Цей процес продовжується, поки всі об’єкти не будуть розподілені по однорідним групам.

У результаті застосування розглянутого методу виходять підмножини, однорідні в змісті ізотропності. Тобто вони розташовані в багатомірному просторі так, що за формою хмара розсіювання більше схожа на кулю, ніж на еліпсоїд.

З погляду потреб економетричного моделювання подібні підмножини являють собою результат штучного, нав'язаного, а не природного розбиття досліджуваної сукупності об'єктів. При такому способі розбивки існує потенційна можливість розділити дійсно однорідні об'єкти, якщо розуміти однорідність як "структурну" подібність об'єктів. Подібна небажана розбивка може виникнути внаслідок того, що в значеннях ознак присутні обидві компоненти (структури і потенціалу) Тому цілком обґрунтованим представляється пропозиція використовувати той спосіб, при якому виділені підмножини будуть ізоморфічно однорідними, – адже тоді вони будуть більш відповідати специфіці економетричних досліджень.

Приклад

Завдання 1.1

Нехай відомі такі дані по 10 підприємствам: X_1 – загальні витрати на одиницю продукції, X_2 – фондовіддача на одиницю продукції. Потрібно розбити сукупність даних на однорідні підмножини. Вихідні дані наведені в табл. 1.1.

Таблиця 1.1

Підприємство, n_i	Загальні витрати на одиницю продукції, X_1	Фондовіддача на одиницю продукції, X_2
1	**0,92**	**0,51**
2	**0,72**	**0,59**
3	**0,83**	**1,03**
4	**0,81**	**1,21**
5	**0,82**	**0,63**
6	**0,93**	**0,68**
7	**0,84**	**0,57**
8	**0,89**	**1,52**
9	**0,89**	**1,04**
10	**0,95**	**0,99**

Розв’язок.

А) Ізотропне перетворення.

Стандартизовані дані наведені в табл. 1.2, а матриця відстаней – в табл. 1.3 В тій же таблиці наведене мінімальне значення кожного стовпчика для пошуку радіуса кулі і кількість значень k_i кожного стовпчика, які менші за радіус.

Таблиця 1.2

Стандартизовні дані

n_i	Z_1	Z_2
1	0,86	-1,10
2	-2,02	-0,86
3	-0,43	0,46
4	-0,72	1,00
5	-0,58	-0,74
6	1,01	-0,59
7	-0,29	-0,92
8	0,43	1,93
9	0,43	0,49
10	1,30	0,34

Таблиця 1.3

	Матриця відстаней										
	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9	n_{10}	
n_1	0	2,89	2,03	2,63	1,48	0,53	1,17	3,06	1,65	1,50	
n_2	2,89	0	2,06	2,27	1,45	3,04	1,73	3,71	2,80	3,52	
n_3	2,03	2,06	0	0,61	1,21	1,78	1,39	1,71	0,86	1,73	
n_4	2,63	2,27	0,61	0	1,75	2,35	1,97	1,48	1,26	2,12	
n_5	1,48	1,45	1,21	1,75	0	1,59	0,34	2,86	1,59	2,16	
n_6	0,53	3,04	1,78	2,35	1,59	0	1,34	2,59	1,22	0,97	
n_7	1,17	1,73	1,39	1,97	0,34	1,34	0	2,94	1,58	2,02	
n_8	3,06	3,71	1,71	1,48	2,86	2,59	2,94	0	1,44	1,81	
n_9	1,65	2,80	0,86	1,26	1,59	1,22	1,58	1,44	0	0,88	
n_{10}	1,50	3,52	1,73	2,12	2,16	0,97	2,02	1,81	0,88	0	ρ
min	0,53	1,45	0,61	0,61	0,34	0,53	0,34	1,44	0,86	0,88	**1,45**
k_i	3	1	5	3	3	5	5	2	6	3	

Таким чином, першу однорідну підмножину складуть такі об’єкти:

$\Omega_1 = \{ n_3, n_4, n_6, n_8, n_9, n_{10} \}$.

Вилучимо ці об’єкти з подальшого розгляду. Одержимо таку матрицю (табл. 1.4). З її аналізу видно, що друга підмножина буде містити такі об’єкти

$\Omega_2 = \{ n_1, n_5, n_7 \}$.

Відповідно, залишається один об’єкт, якій складе третю підмножину.

$\Omega_3 = \{ n_2 \}$.

Таблиця 1.4

	n_1	n_2	n_5	n_7
n_1	0	2,89	1,48	1,17
n_2	2,89	0	1,45	1,73
n_5	1,48	1,45	0	0,34
n_7	1,17	1,73	0,34	0
k_i	2	1	2	3

Для аналізу отриманих результатів запишемо найбільше та найменше значення кожної ознаки та середнє значення ознак для кожної однорідної групи об’єктів (табл. 1.5).

Таблиця 1.5

	X_1	X_2
min	0,72	0,51
max	0,95	1,52
Середнє	0,86	0,88
середнє для Ω_1	0,88	1,08
середнє для Ω_2	0,86	0,57

середнє для Ω_3	0,72	0,59

Отже, однорідні групи характеризуються такими властивостями:

Ω_1 Високі витрати на одиницю продукції, висока фондовіддача.

Ω_2 Середні витрати на одиницю продукції, низька фондовіддача.

Ω_3 Низькі витрати на одиницю продукції, низька фондовіддача.

Б) Ізотонічне перетворення.

Стандартизовані дані занесемо до табл. 1.6, а розраховану матрицю відстаней – до табл. 6.1. Знову обчислимо радіус кулі проведемо розбиття множини об'єктів на однорідні підмножини. Перша підмножина буде містити 8 об'єктів. Перетворену після вилучення цих об'єктів матрицю відстаней подамо у табл. 1.8. Неважко переконатись, що друга підмножина буде містити решту об'єктів. Отже, підмножини матимуть такий вигляд:

$\Omega_1 = \{ n_1, n_2, n_3, n_5, n_6, n_7, n_9, n_{10} \}$.

$\Omega_2 = \{ n_4, n_8 \}$.

Для аналізу результатів запишемо табл. 1.9, аналогічну табл. 1.1.

Отже, за рівнями значень показників об'єкти розподілились так :

Ω_1 – середні витрати на одиницю продукції, низька фондовіддача

Ω_2 – середні витрати на одиницю продукції, висока фондовіддача

Зауважимо, що другий показник (X_2) має більшу варіацію, тому здійснює більший вплив на розподіл об'єктів по множинам.

Таблиця 1.6

Стандартизовані дані			
n_i	U_1	U_2	$W = U_1+U_2$
1	0,107	0,058	0,165
2	0,084	0,067	0,151
3	0,097	0,117	0,214
4	0,094	0,138	0,232
5	0,095	0,072	0,167
6	0,108	0,078	0,186
7	0,098	0,065	0,163
8	0,103	0,173	0,277
9	0,103	0,119	0,222
10	0,110	0,113	0,223

Таблиця 1.7

	Матриця відстаней $d = \| w_i - w_j \|$										
	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9	n_{10}	
n_1	0	0,014	0,049	0,067	0,002	0,021	0,002	0,112	0,057	0,058	
n_2	0,014	0	0,063	0,081	0,016	0,035	0,012	0,126	0,071	0,072	
n_3	0,049	0,063	0	0,018	0,047	0,028	0,051	0,063	0,008	0,009	
n_4	0,067	0,081	0,018	0	0,065	0,046	0,069	0,045	0,010	0,009	
n_5	0,002	0,016	0,047	0,065	0	0,018	0,005	0,110	0,055	0,056	
n_6	0,021	0,035	0,028	0,046	0,018	0	0,023	0,091	0,036	0,038	
n_7	0,002	0,012	0,051	0,069	0,005	0,023	0	0,114	0,059	0,061	
n_8	0,112	0,126	0,063	0,045	0,110	0,091	0,114	0	0,055	0,053	
n_9	0,057	0,071	0,008	0,010	0,055	0,036	0,059	0,055	0	0,001	
n_{10}	0,058	0,072	0,009	0,009	0,056	0,038	0,061	0,053	0,001	0	ρ
min	0,002	0,012	0,008	0,009	0,002	0,018	0,002	0,045	0,001	0,001	0,045
k_i	5	5	5	4	5	8	5	1	5	5	

Таблиця 1.8

	n_4	n_8
n_4	0	0,01

n_8	0,01	0
k_i	2	2

Таблиця 1.9

	X_1	X_2
min	0,72	0,51
max	0,95	1,52
Середнє	0,86	0,88
середнє для Ω_1	0,86	0,76
середнє для Ω_2	0,85	1,37

В) Ізоморфічне перетворення.

Проведемо стандартизацію даних і запишемо їх до табл. 1.10, а матрицю відстаней – в табл. 1.11. Помітимо, що при відборі першої підмножини маємо дві групи об'єктів, які відрізняються складом: по два об'єкти в них різні. В такому випадку проаналізуємо відстані між різними об'єктами. Оскільки $d_{9,10} < d_{2,3}$, то слід зупинитись на другому варіанті підмножини. Отже, перша підмножина містить такі об'єкти:

$\Omega_1 = \{ n_1, n_5, n_6, n_7, n_9, n_{10}\}$.

Перетворена матриця відстаней наведена в табл. 1.12. Як видно, об'єкти розподіляться по двом групам. Отже, в результаті маємо 3 однорідні з точки зору структури даних підмножини

$\Omega_1 = \{ n_1, n_5, n_6, n_7, n_9, n_{10}\}$.

$\Omega_2 = \{ n_3, n_4, n_8 \}$;

$\Omega_3 = \{ n_2\}$.

Таблиця 1.10

Стандартизовні дані

n_i	U_1	U_2	$W = U_1+U_2$	$Z_1=U_1/W$	$Z_2=U_2/W$
1	0,107	0,058	0,165	0,648	0,090
2	0,084	0,067	0,151	0,507	0,104
3	0,097	0,117	0,214	0,584	0,181
4	0,094	0,138	0,232	0,570	0,213
5	0,095	0,072	0,167	0,577	0,111
6	0,108	0,078	0,186	0,655	0,120
7	0,098	0,065	0,163	0,592	0,100
8	0,103	0,173	0,277	0,627	0,268
9	0,103	0,119	0,222	0,627	0,183
10	0,110	0,113	0,223	0,669	0,174

Таблиця 1.11

	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9	n_{10}	
n_1	0	0,142	0,111	0,146	0,074	0,031	0,057	0,179	0,096	0,087	
n_2	0,142	0,000	0,110	0,126	0,071	0,149	0,085	0,203	0,144	0,177	
n_3	0,111	0,110	0,000	0,035	0,071	0,094	0,081	0,096	0,042	0,085	
n_4	0,146	0,126	0,035	0,000	0,102	0,126	0,115	0,078	0,064	0,106	
n_5	0,074	0,071	0,071	0,102	0,000	0,078	0,018	0,164	0,087	0,111	
n_6	0,031	0,149	0,094	0,126	0,078	0,000	0,066	0,151	0,069	0,056	
n_7	0,057	0,085	0,081	0,115	0,018	0,066	0,000	0,171	0,090	0,107	
n_8	0,179	0,203	0,096	0,078	0,164	0,151	0,171	0,000	0,084	0,102	
n_9	0,096	0,144	0,042	0,064	0,087	0,069	0,090	0,084	0,000	0,043	
n_{10}	0,087	0,177	0,085	0,106	0,111	0,056	0,107	0,102	0,043	0,000	ρ
min	0,031	0,071	0,035	0,035	0,018	0,031	0,018	0,078	0,042	0,043	0,078
k_i	4	2	4	3	6	6	4	1	5	3	

Таблиця 1.12

	n_2	n_3	n_4	n_8
n_2	0	0,110	0,126	0,203

n_3	0,110	0	0,035	0,096
n_4	0,126	0,035	0	0,078
n_8	0,203	0,096	0,078	0
k_i	1	2	3	1

Проаналізуємо розподіл об'єктів в такому випадку. Для цього визначимо структуру ознак для кожного об'єкта, поділивши значення ознак. Результати для аналізу наведені в табл. 1.13.

Як видно, першу групу переважно складають об'єкти, в яких значення частки ознаки X_1 більше значення частки ознаки X_2. Відмінність спостерігається лише останніх двох об'єктів. Це свідчить про недолік методу куль для ізоморфічних перетворень – в однорідну підмножину увійшли об'єкти з іншою структурою даних. Друга група включає об'єкти, для яких частка другої ознаки переважає частку першої. Третя підмножина містить об'єкти з приблизно однаковими частками ознак.

Таблиця 1.13

Структура даних		
X_1	X_2	X_1/X_2
0,92	**0,51**	1,804
0,72	**0,59**	1,220
0,83	**1,03**	0,806
0,81	**1,21**	0,669
0,82	**0,63**	1,302
0,93	**0,68**	1,368
0,84	**0,57**	1,474
0,89	**1,52**	0,586
0,89	**1,04**	0,856
0,95	**0,99**	0,960

3. Питання для самоперевірки.

1. Які є види однорідності підмножин? Чим вони характеризуються?
2. Як провести ізотропне перетворення вихідних даних?
3. Як провести ізотонічне перетворення вихідних даних?
4. Як провести ізоморфічне перетворення вихідних даних?
5. В чому сутність методу куль? Які його недоліки?